Regression and Cross-Validation

ABSTRACT

PURPOSE: Implement linear regression to analyze the relationship between short-term instability(fragility index) and male population across age groups, and evaluate the reliability of a given age group's population as an indicator of country fragility.

METHODS: The given dataset “fragility2013male.csv” was loaded into MATLAB using instructor-provided code. The program standardized the data and then used linear regression to find the RMS errors, treating each column(age group) as a vector of the dependent variable fragility index. Additionally, it used the index of the age group with the smallest positive correlation RMS error and performed 5-fold cross-validation to explore its reliability as a proxy for fragility indexes.

RESULTS: Running the program produced a candidate age group to be used as a proxy for the fragility index. The 5-fold cross-validation results indicated minimal RMS errors for both training and testing with consistent values with little variation between train and testing indicating a reliable relationship male population and fragility index

CONCLUSIONS: The use of linear regression to determine the relationship between the fragility index and the male population between age groups has proven to be useful with the provided data set. However, the real-world applicability of this relationship is questionable as there are a myriad of factors that are unaccounted for as the dataset only compares male age demographics to the stability of an entire country.

INTRODUCTION

SCIENTIFIC PURPOSE/OBJECTIVE: The scientific purpose is to evaluate the effectiveness of linear regression and k-fold cross-validation for using another variable to explain/predict the dependent variable.

BACKGROUND: The technical problem in linear for this assignment is to estimate a weight vector w for a design matrix A and a data vector c. The goal is to approximate the weight vector w such that Aw = c. (Ellis) This linear regression is used to identify the age group that contributes most significantly to the dependent variable (fragility index) by checking the RMS error of each group. To assess the performance of the linear model we use k-folds cross-validation. This involves dividing the original dataset M into k smaller subsets, each containing M/k entries. The learning algorithm is trained on the k-1 subsets and tested on the final subset, repeating a total of k times with a different subset used as the test set each time. The average performance of each iteration is used to determine the overall performance of the model. (Arya) The program also standardizes the data, standardizing data ensures that all data has a mean of 0 and a standard

deviation of 1, this makes the model less dependent on the scale of individual features while also removing or reducing collinearity. (Towards AI Team)

SCIENTIFIC QUESTION: With the provided dataset, is the method of using linear regression able to provide an independent variable that is capable of acting as a proxy/indicator for the fragility index

This is tested by implementing linear regression and k-fold cross-validation in MATLAB and applying it to the dataset. Its effectiveness is evaluated through observation of the generated RMS values from k-fold cross-validation.

METHODS

The initial step is to load and preprocess the dataset, this is done by loading the dataset using the MATLAB's builtin function. The data is then separated into various parts and the content is evaluated for its size.

With the data separated into FragilityIndex and dataMatrix, the program standardizes and normalizes the data to improve consistency. No intercept term was implemented as the nature of the data was considered and a male population of 0 for a given country (If this were true the process used to obtain a country's fragility probably does not apply)

Using the standardized data the RMS errors of linear regression of the data were found by treating each column individually and calculating the weight vector for the standardized data by dividing it by the dependent variable.(fragility) The program also calculates the predicted dependent variable by multiplying the weight vector by the independent variable. The RMS error was then calculated by utilizing the prebuilt function from MATLAB on the result of subtracting the dependent variable by the predicted dependent variable. Correlation values were also obtained by utilizing the linsolve builtin function.

With this process done on each independent variable, the indexes of the smallest RMS error of the positive correlations and the negative correlations were obtained. Utilizing the positive correlation RMS 5-fold cross-validation was performed. To do this we first randomly shuffled the data to remove ordering and selection bias, then, subsets of the data were extracted to create test and training sets which change with each fold. The folds were performed using the implemented mykfold function which changes the testing and training data following the k-fold cross-validation technique. By using this technique, the assessment of the regression is improved as all data points are used in both testing and training in multiple combinations.

RESULTS

Table 1: RMS Errors in standardized units.

Age Groups	RMS Error
0-4	72.0449
5-9	72.0214
10-14	72.0419
15-19	72.1961
20-24	73.1887
25-29	74.2650
30-34	74.1894
35-39	73.5094
40-44	72.6477
45-49	72.1774
50-54	72.2344
55-59	72.2397
60-64	72.2105
65-69	72.0631
70-74	72.2366
75-79	72.2990
80-84	72.1148
85-89	72.0165
90-94	72.1190
95-99	72.6991
100+	73.2860

Table 2: Variable(Index 2) best suited to be a proxy for Fragility Index.
Corresponding age group 5-9

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
RMS Train	14.6997	14.5025	14.2773	15.4654	14.8013
RMS Test	15.1674	15.9826	16.7702	11.7990	14.7331

DISCUSSION

The results of the linear regression model showed that the age group 5-9 is best suited to act as a proxy Fragility Index (dependent variable), based on the analysis using 5-fold cross-validation to validate the results.

No intercept term was used in the linear regression analysis as it was decided that the population of 0 males in any age category would be far removed from ordinary circumstances and should that be the case. The Fragility Index as an indicator of a given country's short-term instability would be ineffective or completely irrelevant.

The choice to standardize the data was made to improve consistency, remove collinearity and remove the impact of scale. The initial observation of the data indicated that there was a large difference in scale compared to the fragility index and population numbers. Before standardization, there were no negative correlation values to be found as any value in population dwarfed the fragility index in comparison. The data was standardized using MATLAB's built-in zscore function.

Looking at the results of the 5-fold cross-validation, the selected age group 5-9 shows a strong relationship to the Fragility index. With an average RMS error of 14.89046, it indicates a decently high amount of error. It should be of note that testing data error seems to be above training with the exception of fold 4 and fold 5 slightly. Overall the population of young males in a country are only able to be loosely tracked onto the Fragility Index of a country. Furthermore, it should be stressed that it can not be the sole metric used as it lacks large amounts of political, cultural and economic context that could influence a country's level of instability.

REFERENCES

Arya, Nisha. “Why Use K-Fold Cross Validation?” *KDnuggets*, 11 July 2022, www.kdnuggets.com/2022/07/kfold-cross-validation.html.

Ellis, Randy. “Assignment 2.” *Https://Onq.queensu.ca/Content/Enforced/859947-CISC271W24/Homework/A2.Pdf?IsCourseFile=True*, onq.queensu.ca/content/enforced/859947-CISC271W24/homework/A2.pdf?isCourseFile=true. Accessed 6 Feb. 2024.

IBM. “About Linear Regression | IBM.” *Www.ibm.com*, www.ibm.com/topics/linear-regression.

Towards AI Team. “How, When, and Why Should You Normalize / Standardize / Rescale Your Data? – towards AI — the Best of Tech, Science, and Engineering.” *Towardsai*, 16 May 2019, towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.